



NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Annotating the human genome: uses and pitfalls of computers in biology

Dr. Manfred Zorn

March 19, 2001
Old Dominion University
Norfolk, VA



Outline



- **Introduction**
- **Basic Biology**
- **Human Genome Project**
- **After the Genome**

Old Dominion University



Lawrence Berkeley National Laboratory



- **Founded in 1931 by Ernest O. Lawrence**
- **Best known for Particle Physics, found a dozen new transuranic elements: Bk, Cf, Am, Lw, Pu, ..., Sg**
- **About 4000 people, 800 students, 2000 visitors**
- **National User Facilities:**
 - Advanced Light Source
 - NERSC Supercomputing Center

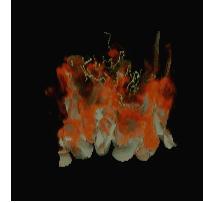
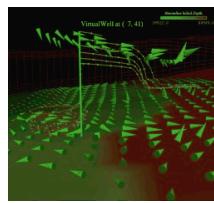
Old Dominion University



NERSC - Overview



- **the Department of Energy, Office of Science, supercomputing facility**
- **unclassified, open facility; serving >2000 users in all DOE mission relevant basic science disciplines**
- **25th anniversary in 1999**



Old Dominion University



NERSC-3 IBM SP



- **NERSC-3
(IBM SP3/RS 6000)**
- **Phase I: June 1999**
 - 608 processors
 - 410 gigaflop peak performance
 - Provides one teraflop NERSC capability
- **Phase II: December 2000**
 - 2,432 processors
 - 3.2 teraflop peak performance
 - 4 teraflop total NERSC capability



Old Dominion University



Center for Bioinformatics and Computational Genomics



- Vision
 - A national center for understanding information and information systems in modern biology
- History
 - Established July 1998 within NERSC at LBNL by merging the Bioinformatics Group and the Human Genome Field Office
 - Co-directed by Sylvia Spengler and Manfred Zorn



Old Dominion University



Center for Bioinformatics and Computational Genomics

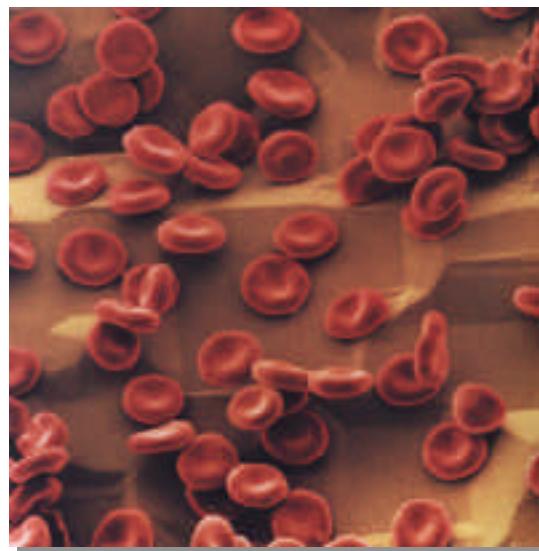


- Research
 - **Special Analysis Tools: Fold Prediction, Phylogeny, genome comparisons**
 - **Compute-intensive Algorithms: clustering, phylogeny**
- Development and Support
 - **Large-scale Genome Annotation**
 - **Wet lab support for Biologists**
- Public Service
 - **Public databases**
 - **Education and Outreach, Standards**

Old Dominion University



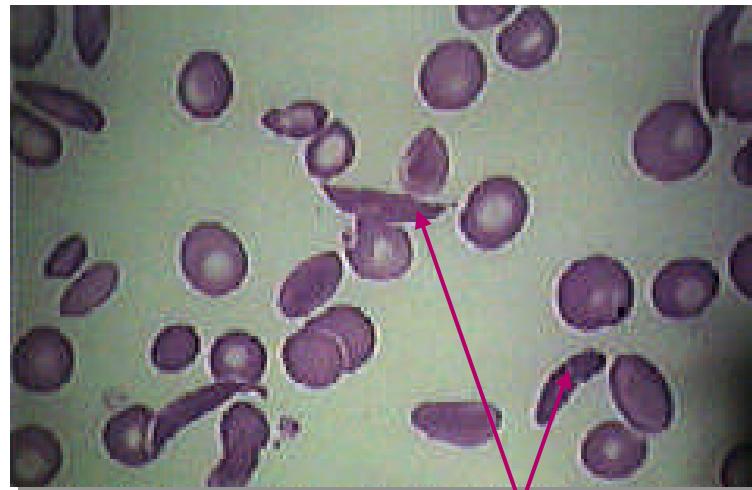
Chocolate Mints?



Old Dominion University



Diagnosis - Blood Smear



Sickle red cells

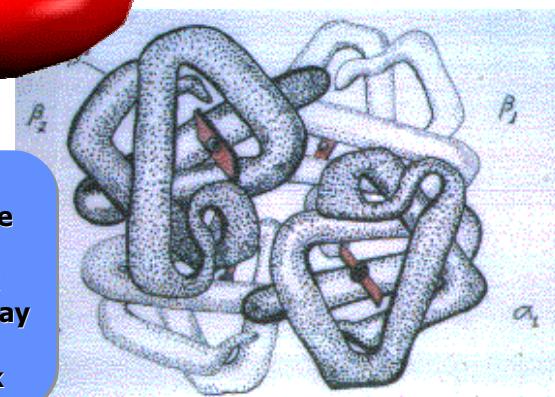
Old Dominion University



Red Blood Cells - Hemoglobin



Hemoglobin is the main chemical in the red blood cell that does all of the work carrying oxygen away from the lungs and carbon dioxide back



Redrawn From Dickerson and Gees, 1964

Old Dominion University



Normal vs. Sickle Hemoglobin



Normal

- disc-Shaped
- soft(like a bag of jelly)
- easily flow through small blood vessels
- lives for 120 days



Sickle

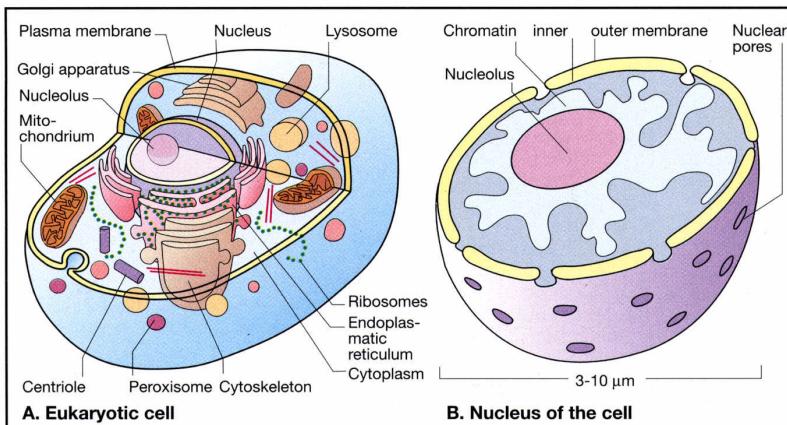
- sickle-Shaped
- hard (like a piece of wood)
- often get stuck in small blood vessels
- lives for 20 days or less



Old Dominion University



Cell Structure



ZBD9806-01631.TIF

Old Dominion University



Truth and Conventional Wisdom in Biology

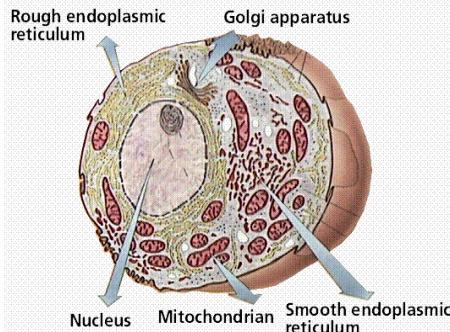


- **Biologists dislike generalizations.**
- **The truth in biology is always more complex than the statement about it.**
- **It is hard to distinguish between fact and fashion in biology.**

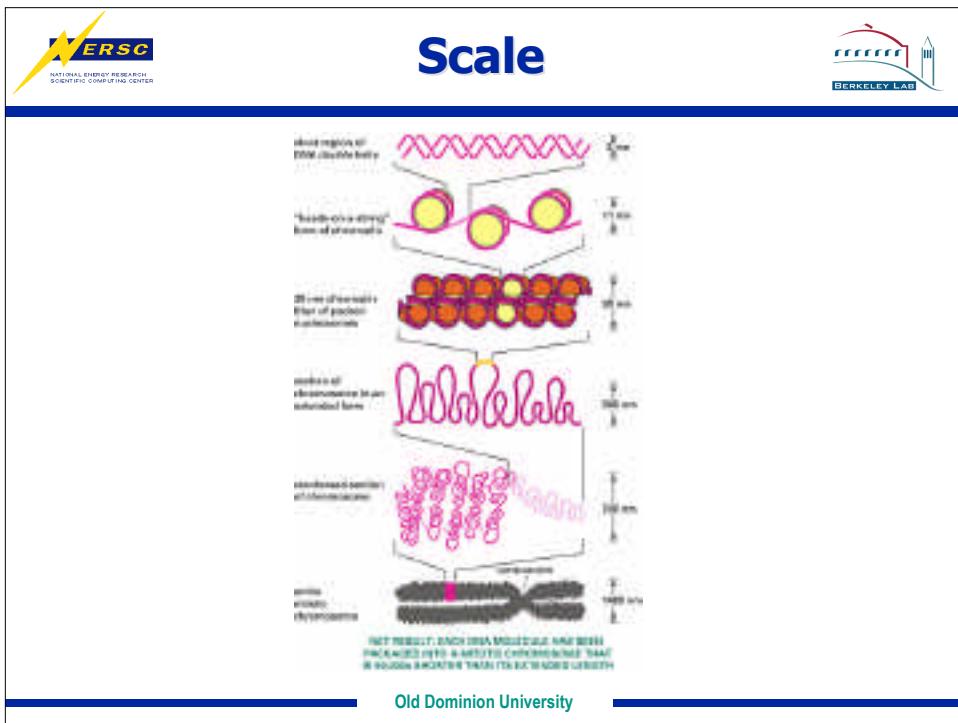
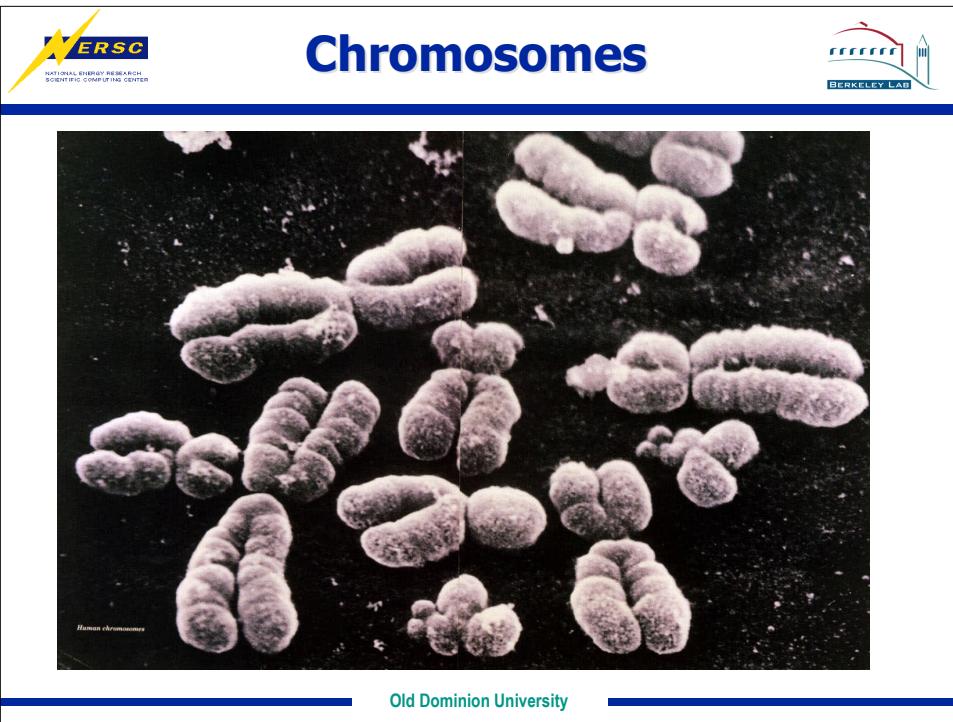
Old Dominion University



Basic Biology



Old Dominion University





The Human Genome



- **24 Chromosomes**

- ✓ 1 – 22, X, Y
- ✓ 23 pairs

- **1 Mitochondrial Genome**

- **3 Billion Base Pairs**

- **~30,000 Genes**

Old Dominion University



How Big?

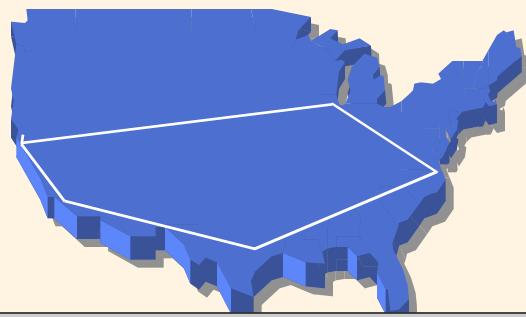


3×10^9 seconds = 95 years

Old Dominion University



One Human Sequence



Typed in 10-pitch font, one human sequence would stretch for more than 5,000 miles. Digitally formatted, it could be stored on one CD-ROM. Biologically encoded, it fits easily within a single cell.

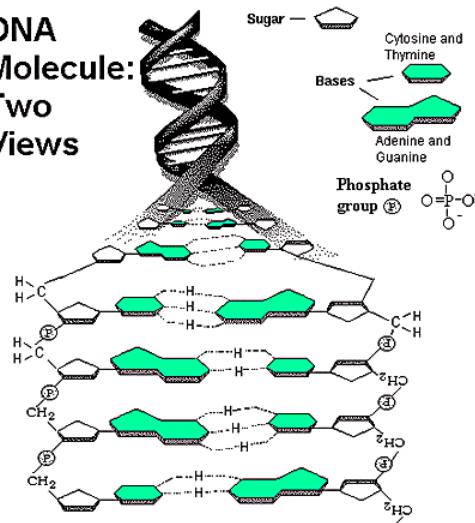
Old Dominion University



DNA - Two Views



DNA Molecule: Two Views



Old Dominion University

Replication

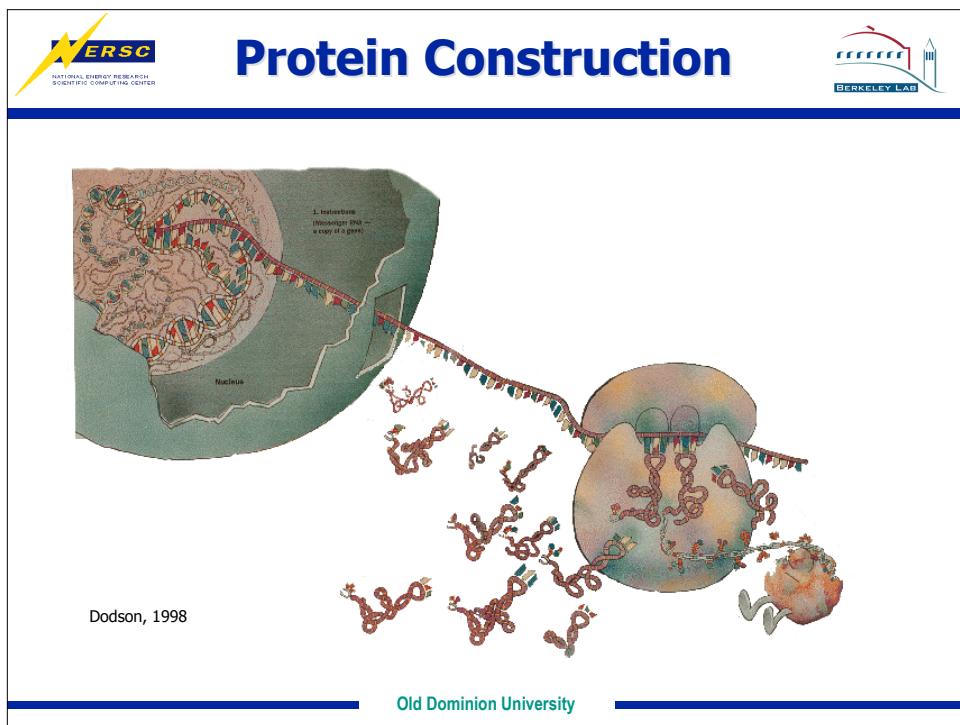
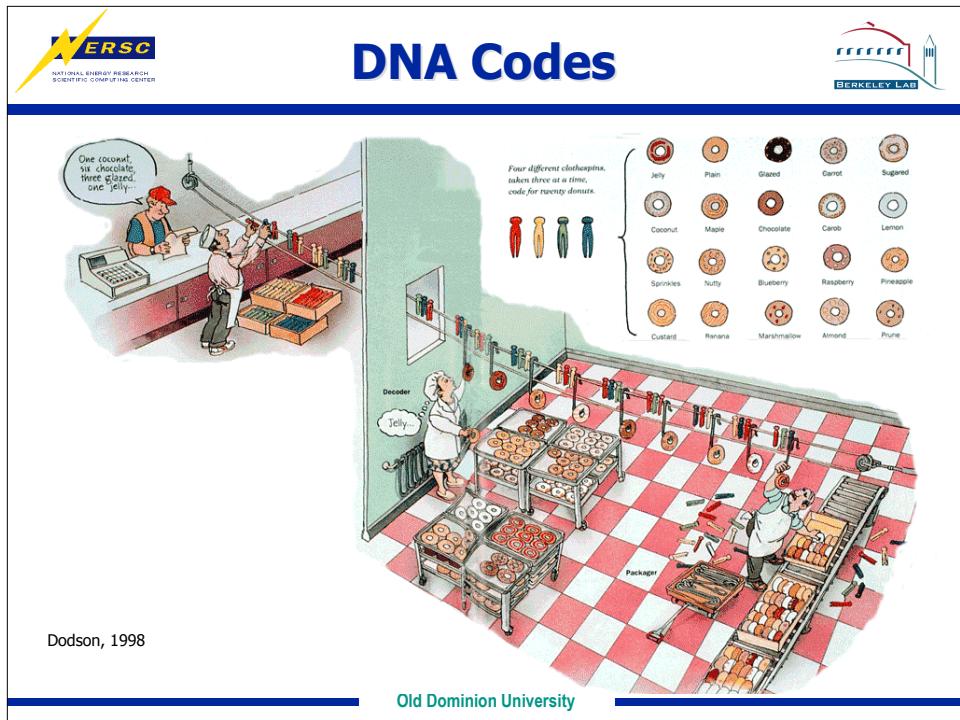
Old Dominion University

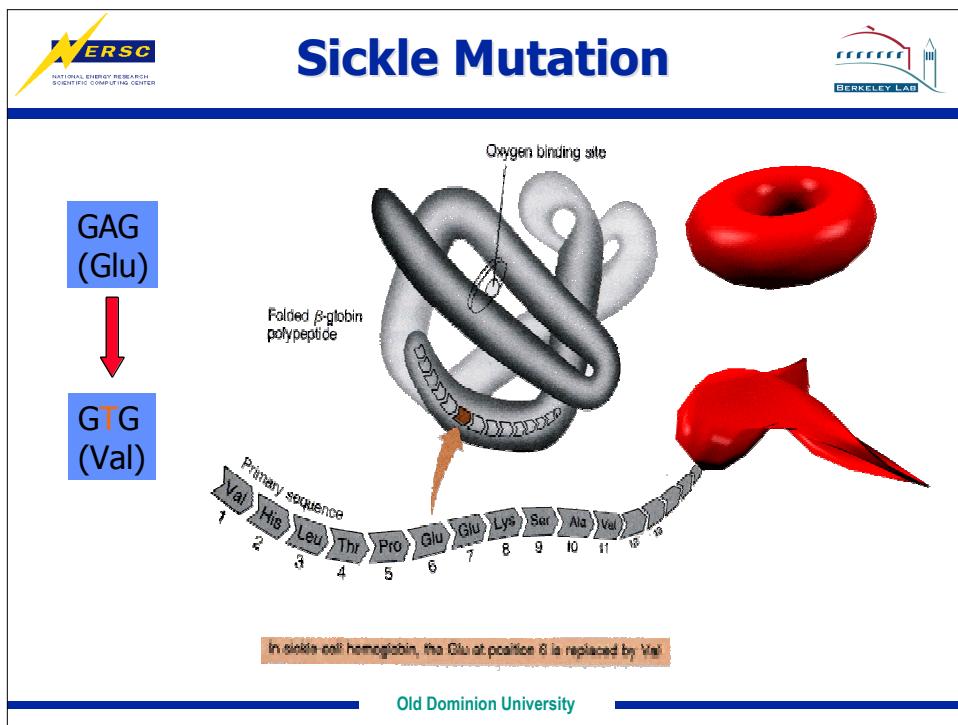
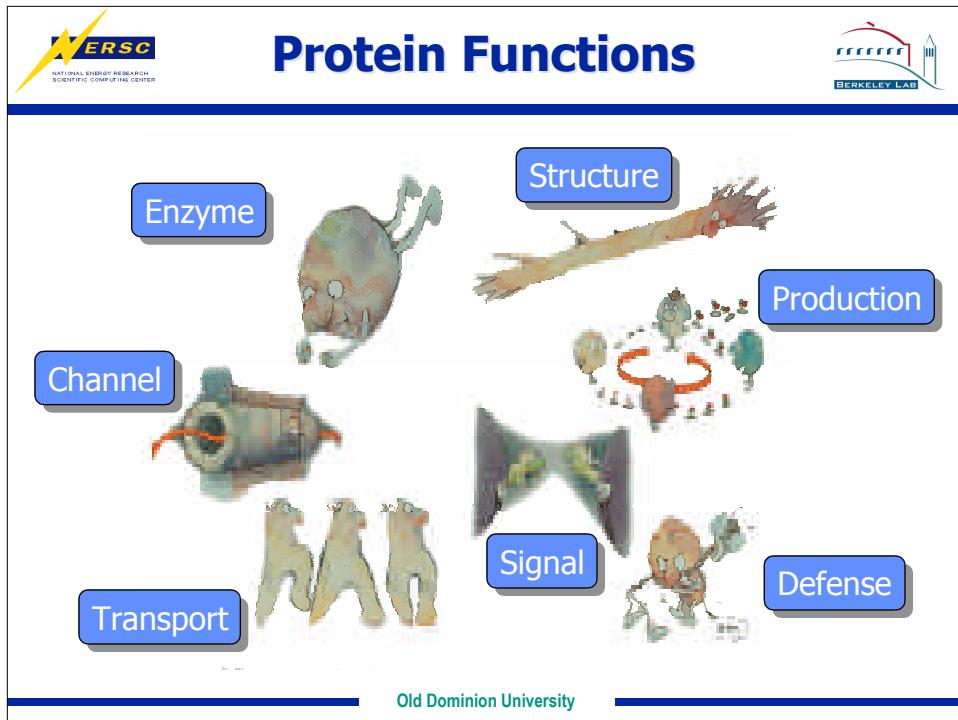
Central Dogma

The fundamental dogma of molecular biology is that genes act to create phenotypes through a flow of information from DNA to RNA to proteins, to interactions among proteins (regulatory circuits and metabolic pathways), and ultimately to phenotypes. Collections of individual phenotypes constitute a population.

The Central Dogma of Molecular Biology

Old Dominion University







Is this a disease?



Normal

- disc-Shaped
- soft(like a bag of jelly)
- easily flow through small blood vessels
- lives for 120 days



Sickle

- sickle-Shaped
- hard (like a piece of wood)
- often get stuck in small blood vessels
- lives for 20 days or less



Well, it depends!

Old Dominion University



Genetic inheritance



Normal/normal



Sickle/normal



- reduced oxygen transport capability
- normal life expectancy
- resistant to malaria infection

Normal/sickle



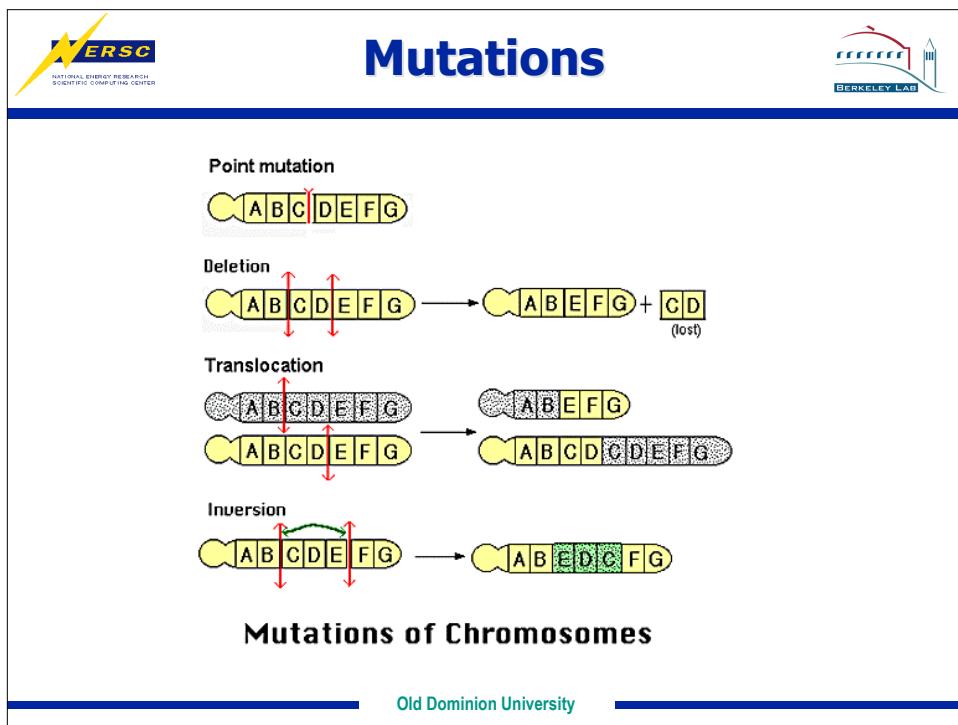
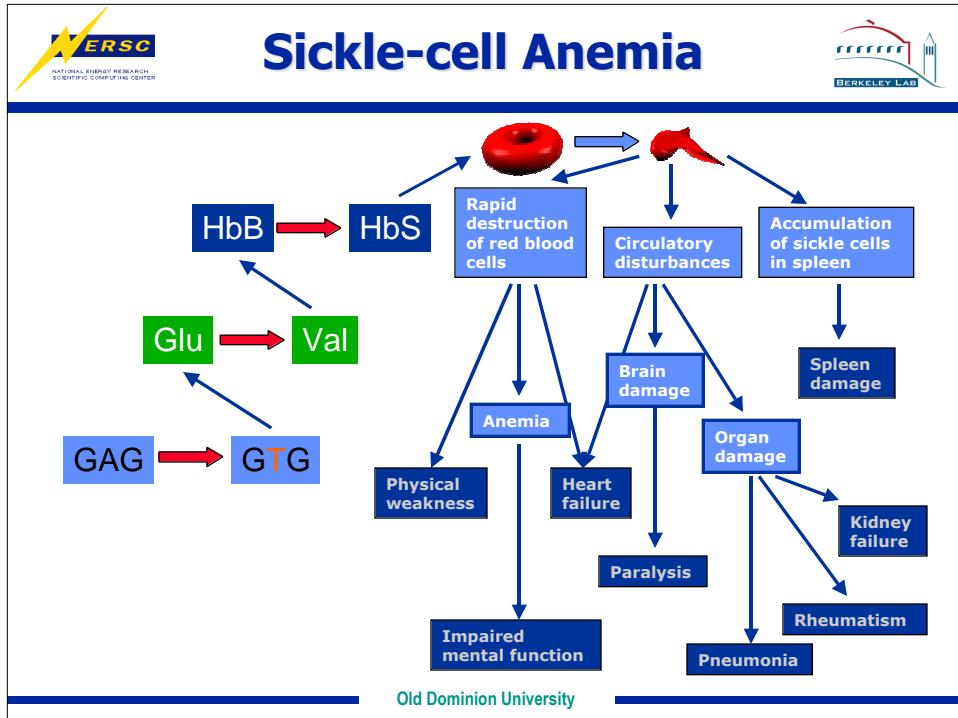
- reduced oxygen transport capability
- normal life expectancy
- resistant to malaria infection

Sickle/sickle



- impaired oxygen transport
- reduced life expectancy

Old Dominion University





Small Changes BIG Effects



- DNA modifications can have a big effect
- Radiation causes DNA damage

**US Department of Energy's
Long-standing program in radiation biology**

Old Dominion University



Genome Project Timeline



- 1984
 - ✓ Department of Energy and Intl. Commission on Protection Against Environmental Mutagens and Carcinogens in Alta, Utah.
- 1986
 - ✓ DOE announces Human Genome Initiative
- 1987
 - ✓ NIH Director establishes Office of Genome Research
- 1988
 - ✓ NRC Mapping and Sequencing the Human Genome
 - ✓ Berkeley Lab launches Human Genome Center
- 1990 Human Genome I

Old Dominion University



Genome Timeline cont'd



■ September 1994

- ✓ First complete map of all human chromosomes one year ahead of schedule.

■ May 1995

- ✓ First genome sequenced: *H. influenzae*

■ May 1998

- ✓ Celera announces commercial project
- ✓ Public effort regroups to five major centers

■ June 2000

- ✓ Joint announcement by NHGRI – Celera

■ February 2001

- ✓ Publication of **We're done!**

Old Dominion University



Genome Projects



1995 <i>H. influenzae</i>	2 Mb
1996 <i>S. cerevisiae</i>	12 Mb
1997 <i>E. coli</i>	5 Mb
1998 <i>C. elegans</i>	100 Mb
1999 Human Chromosome 22	34 Mb
2000 <i>D. melanogaster</i>	140 Mb
2000 <i>H. sapiens</i>	3,000 Mb

Old Dominion University



DNA Sequencing



Read base code from storage medium!

- **Read length: About 600 bases at once**
- **Reader capacity**
 - ✓ 100 lanes in parallel in about 5 hours
 - ✓ 100 lanes in parallel in about 2 hours

3 Billion year old program store

Old Dominion University



NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

DNA Analysis

BERKELEY LAB

Disassemble the base code!

- **Find the genes**
 - Heuristic signals
 - Inherent features
 - Intelligent methods
- **Characterize each gene**
 - Compare with other genes
 - Find functional components
 - Predict features

Old Dominion University

NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

What is a Gene?

BERKELEY LAB

University of Pennsylvania
Computational Biology and
Informatics Laboratory

A. L. Iyer & C. D. Walsh

Old Dominion University



Heuristic Signals



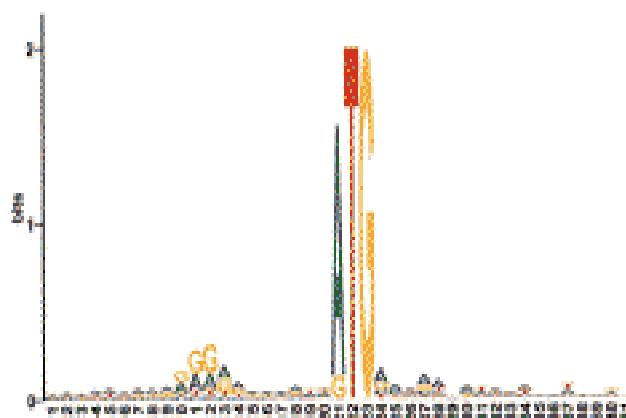
**DNA contains various recognition sites
for internal machinery**

- Promoter signals
- Transcription start signals
- Start Codon
- Exon, Intron boundaries
- Transcription termination signals

Old Dominion University



Start Codon



Old Dominion University

The slide features the NERSC logo at the top left and the Berkeley Lab logo at the top right. The title 'Heuristic Signals' is prominently displayed in large blue letters. A blue banner across the middle contains the text 'Start of the gene'. The background is filled with a dense, illegible sequence of characters, likely representing genetic data.



Inherent Features



DNA exhibits certain biases that can be exploited to locate coding regions

- Uneven distribution of bases
- Codon bias
- CpG islands
- In-phase words
- Encoded amino acid sequence
- Imperfect periodicity
- Other global patterns

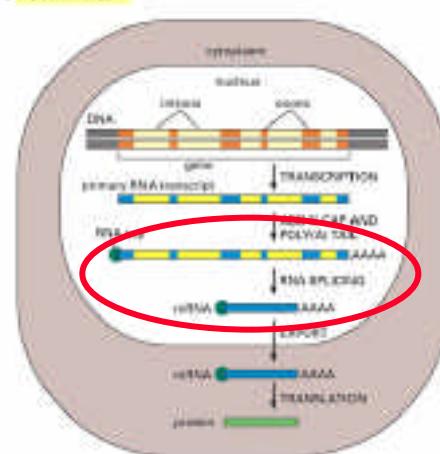
Old Dominion University



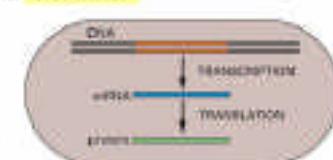
Translation



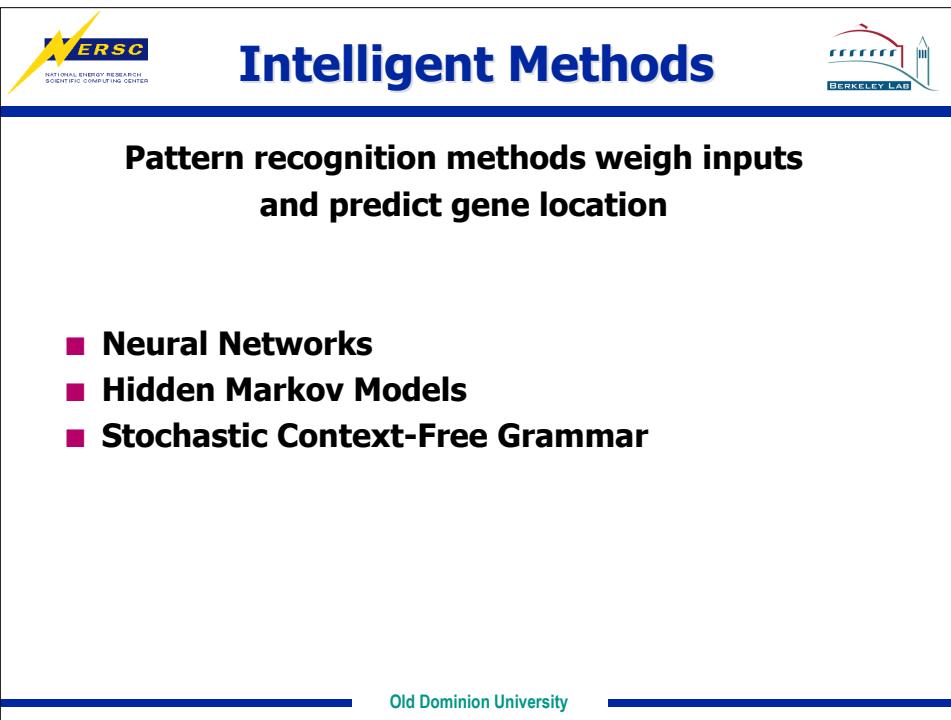
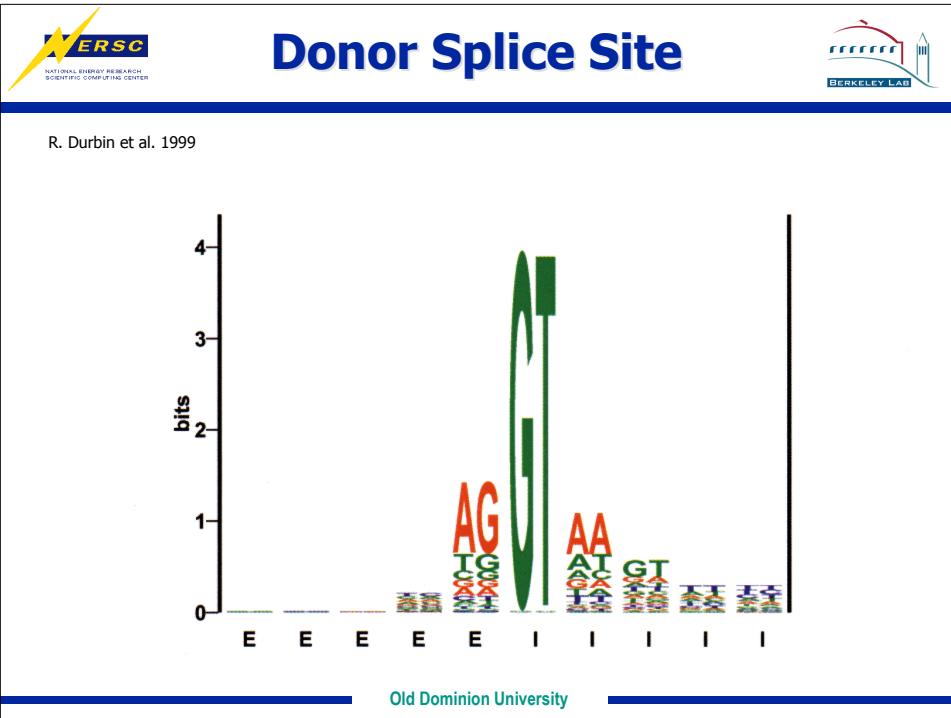
(a) EUKARYOTES



(b) PROKARYOTES



Old Dominion University

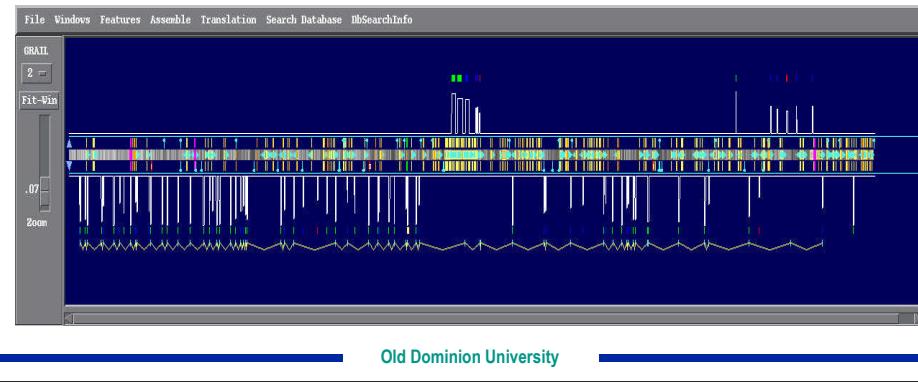




Analyzing Complex Multi-Gene Regions



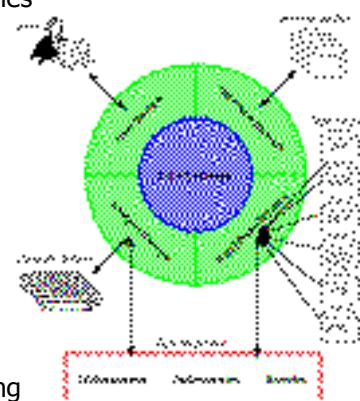
- Errors in exon prediction and splice site boundaries
- Gene boundaries uncertain
- Genes can be on both strands



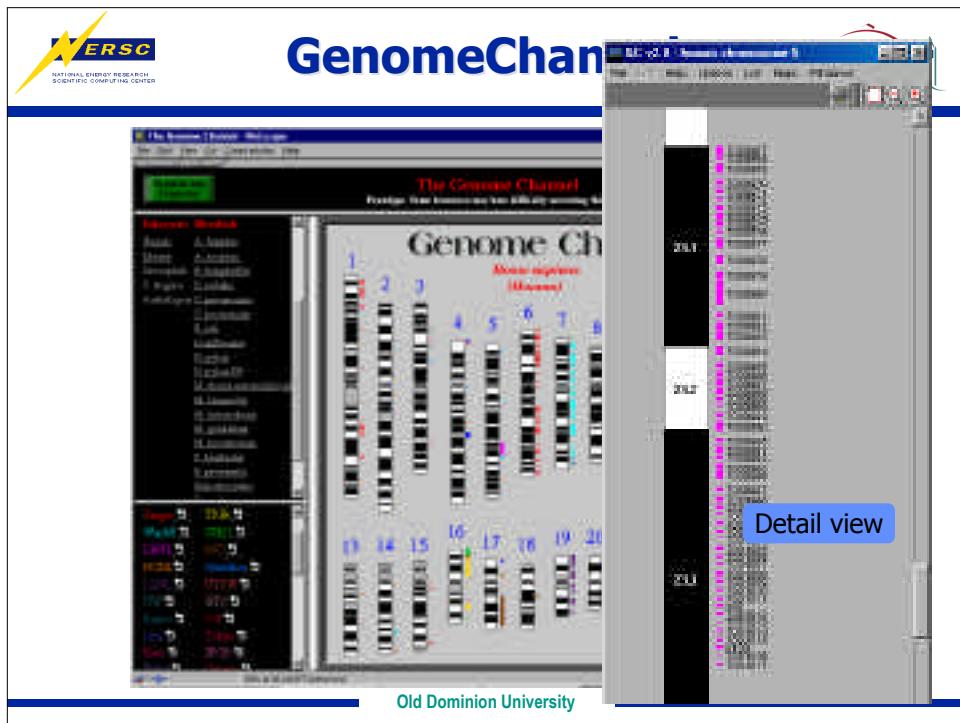
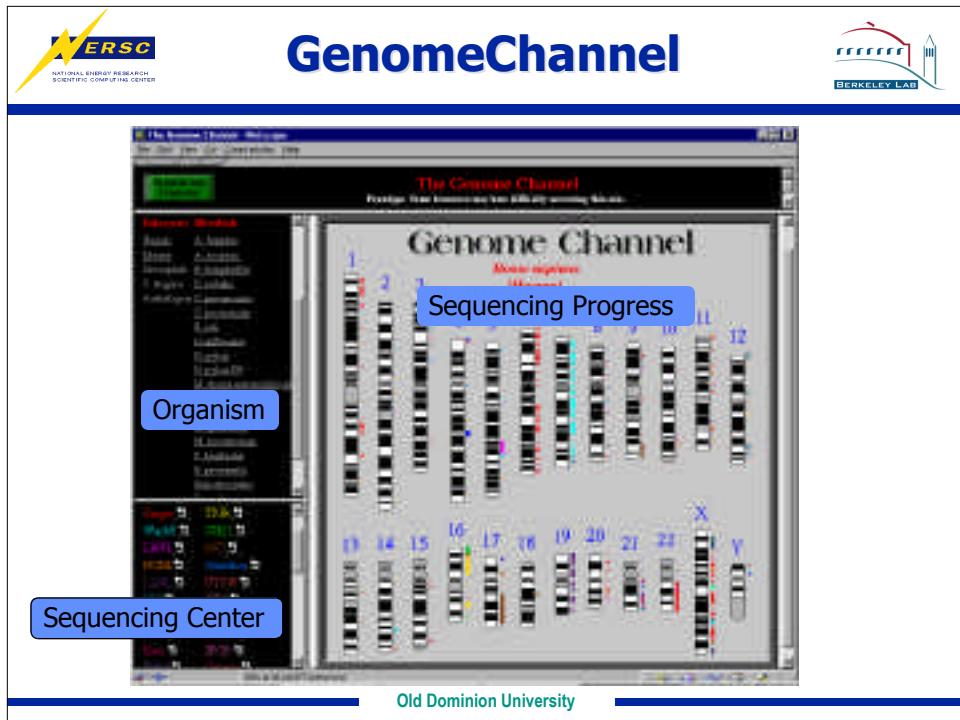
Large-scale Genome Annotation

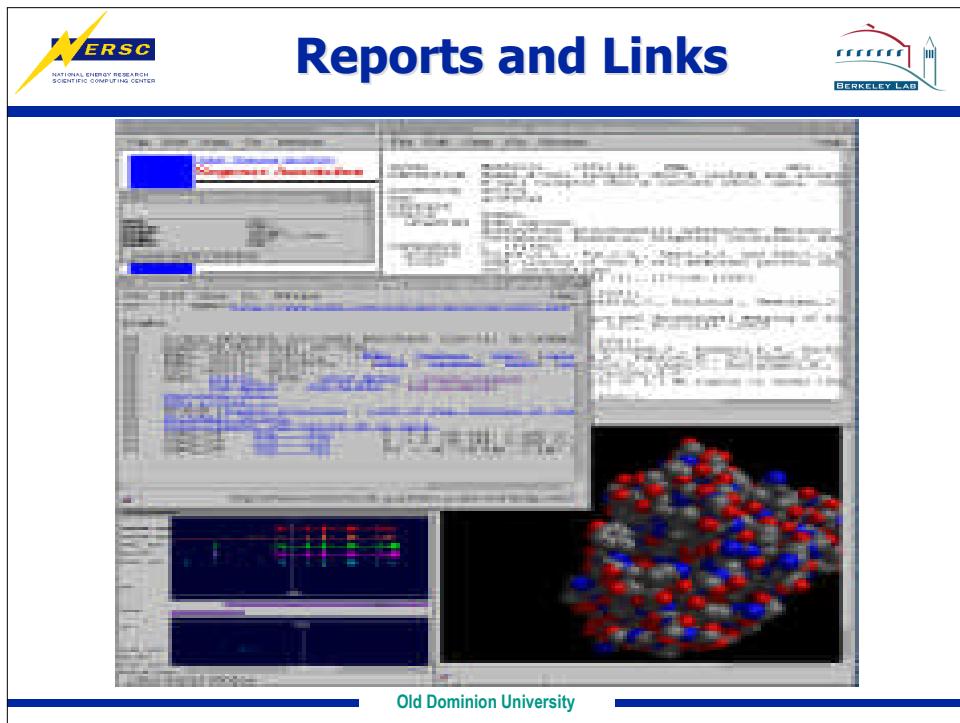
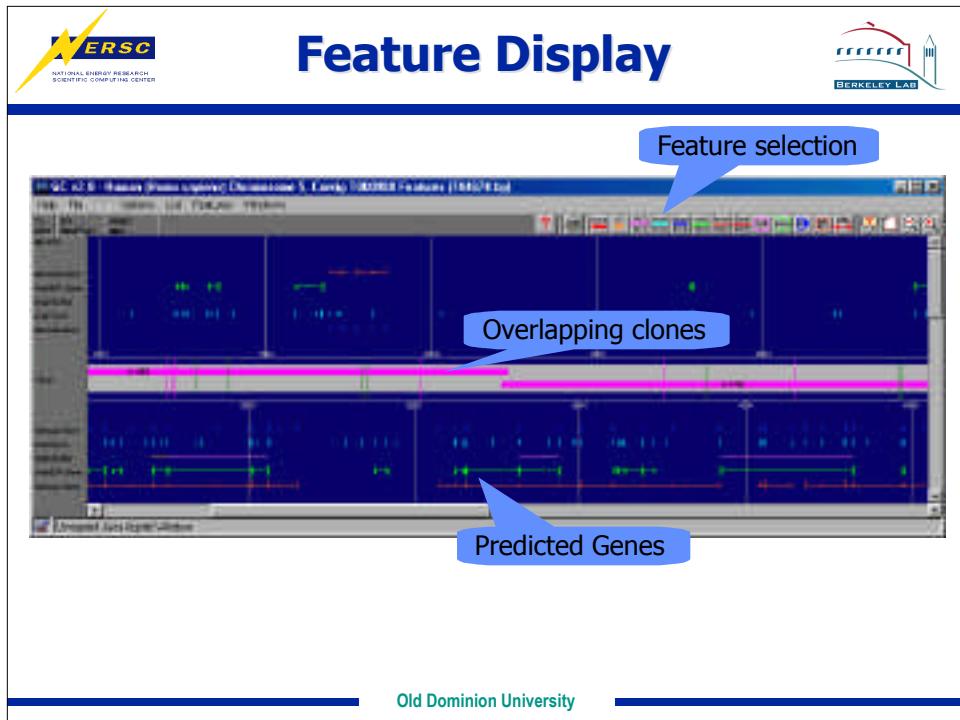


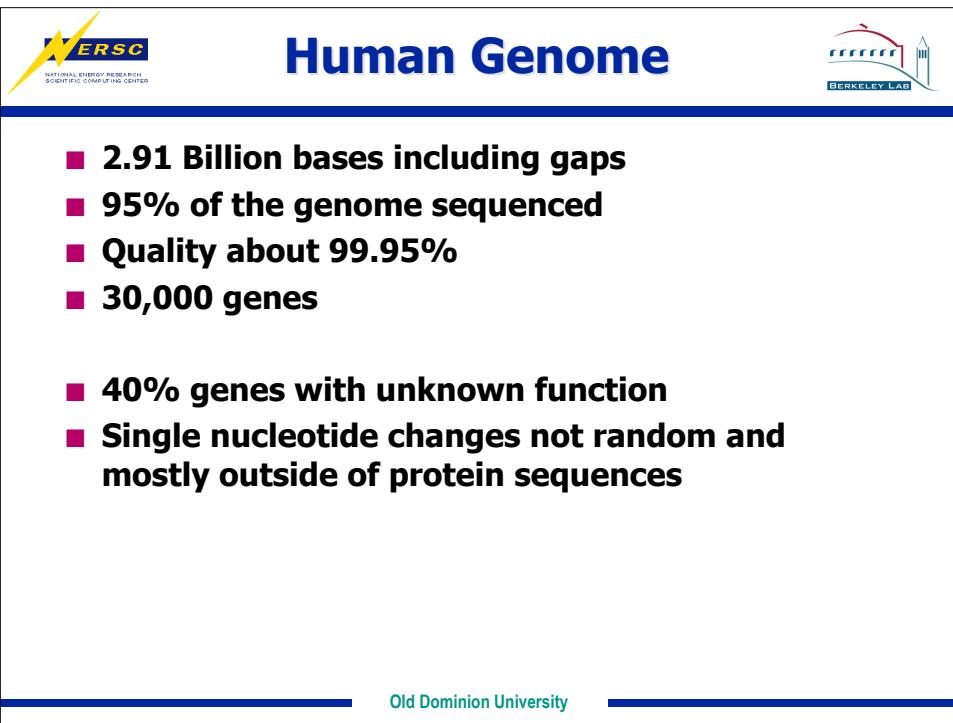
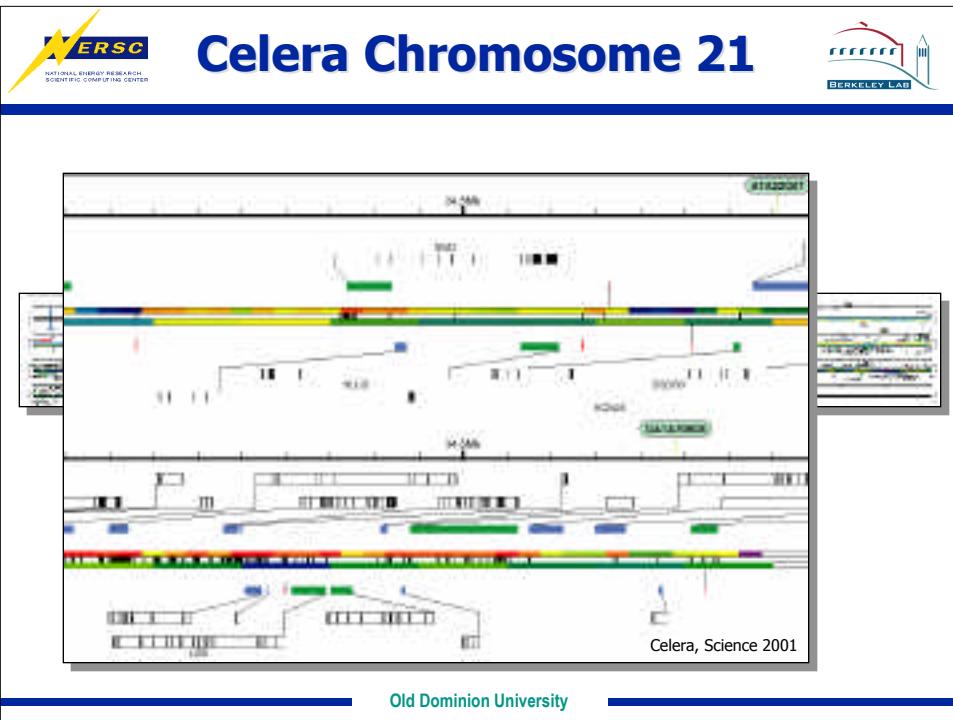
- Multi-laboratory Project
- Standard Annotation of Genomes
 - Genome Channel
 - Genome Catalog
- Comprehensive integration of
 - Analysis tools
 - Data management systems
 - Data mining
 - User services
- Extensible Framework
 - High-performance computing
 - Data integration technology
 - Artificial intelligence



Old Dominion University









Post-Genomics



- **Flood of data**
- **Integration of diverse information**
- **Industrialization of biology**

- **Integration of traditional biology**

- **Quality control**

Biology is an Information Science

Leroy Hood

Old Dominion University



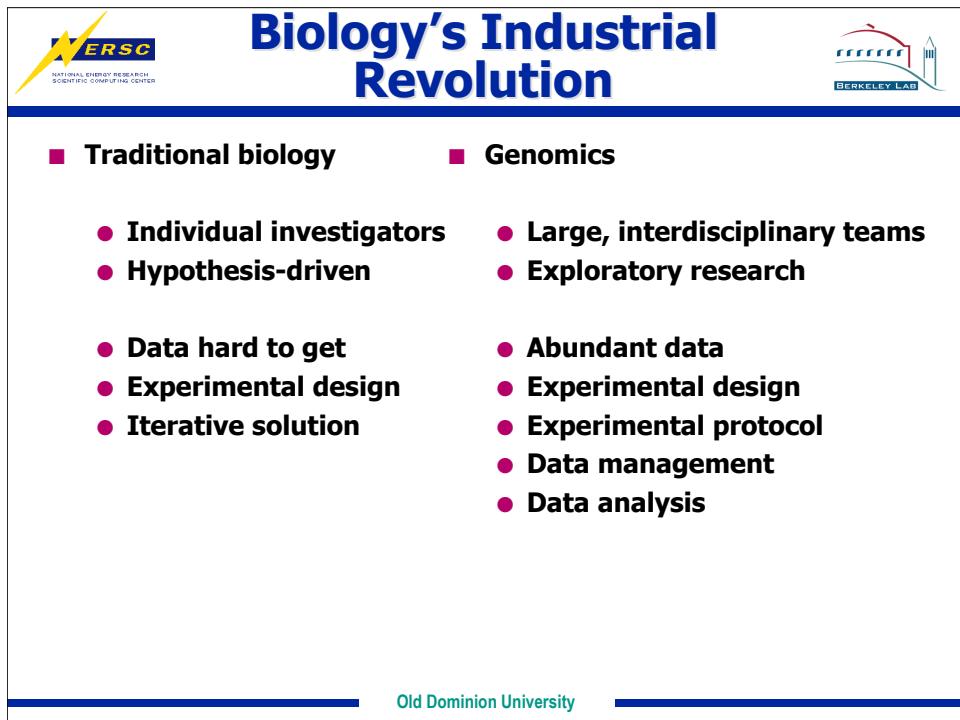
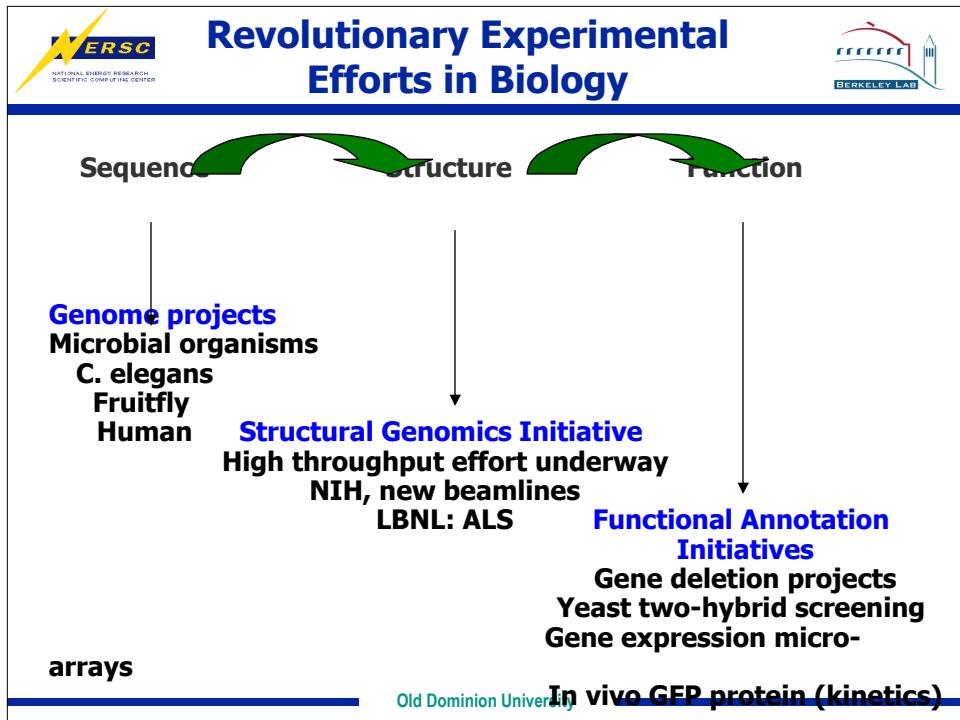
Flood

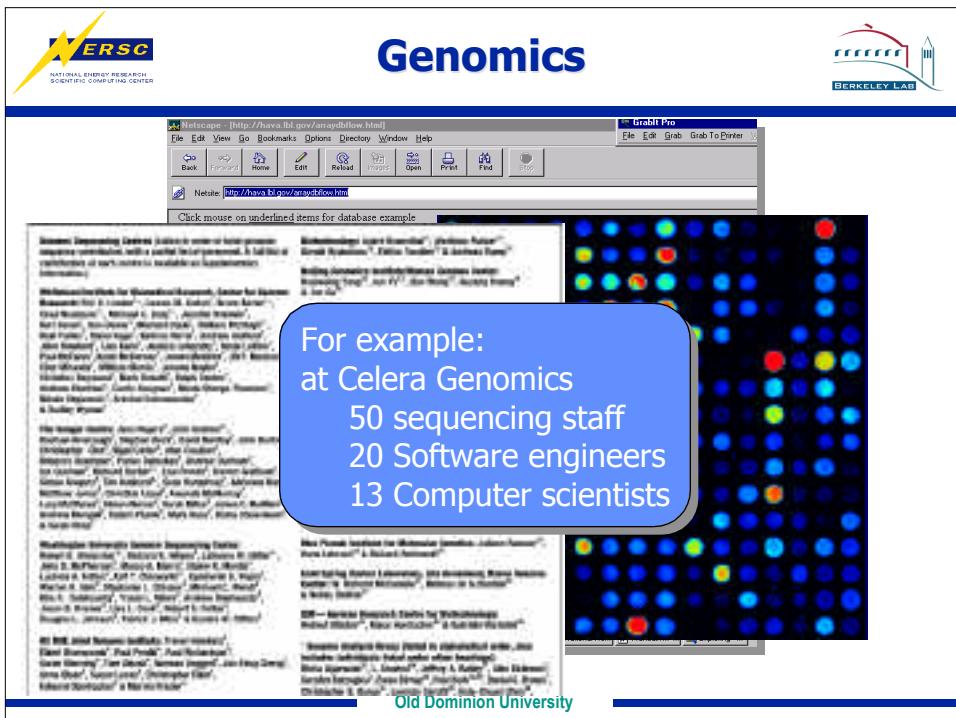
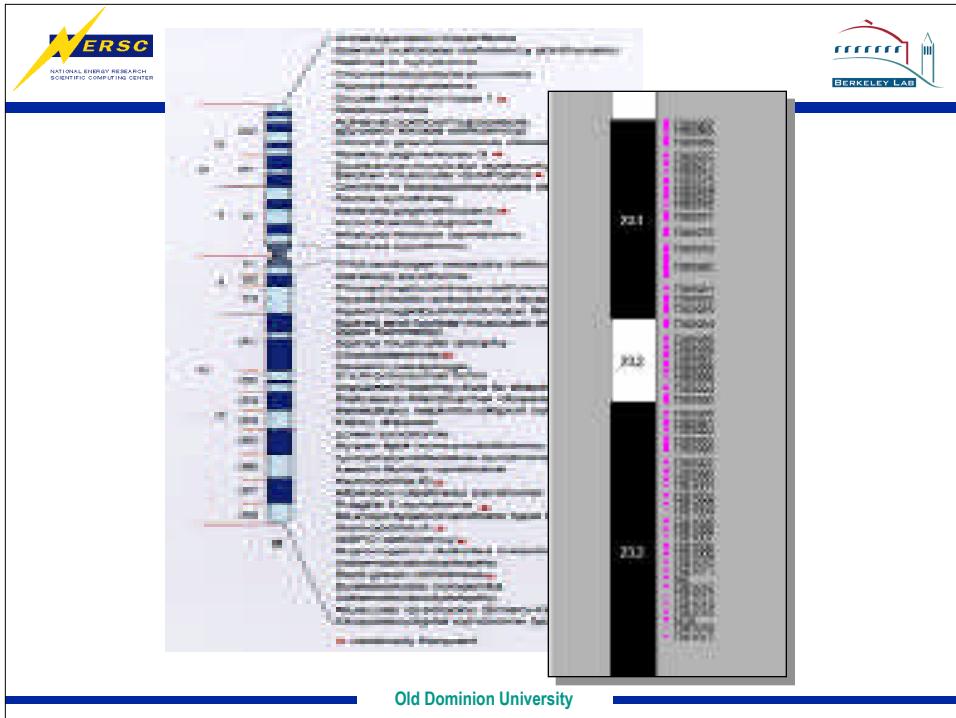


- **The flood of data from genome-wide analysis is transforming biology. We need to develop new, interdisciplinary approaches to convert these data into information about the components and structures of individual biological pathways and to use the resulting information to yield knowledge about general principles that explain the functions and evolution of life.**

Andrew W. Murray, *GenomeBiology* 2000, 1:comment003.1-003.6

Old Dominion University







Caution



Don't get too excited...

...Nature has a few more tricks in store.

Old Dominion University



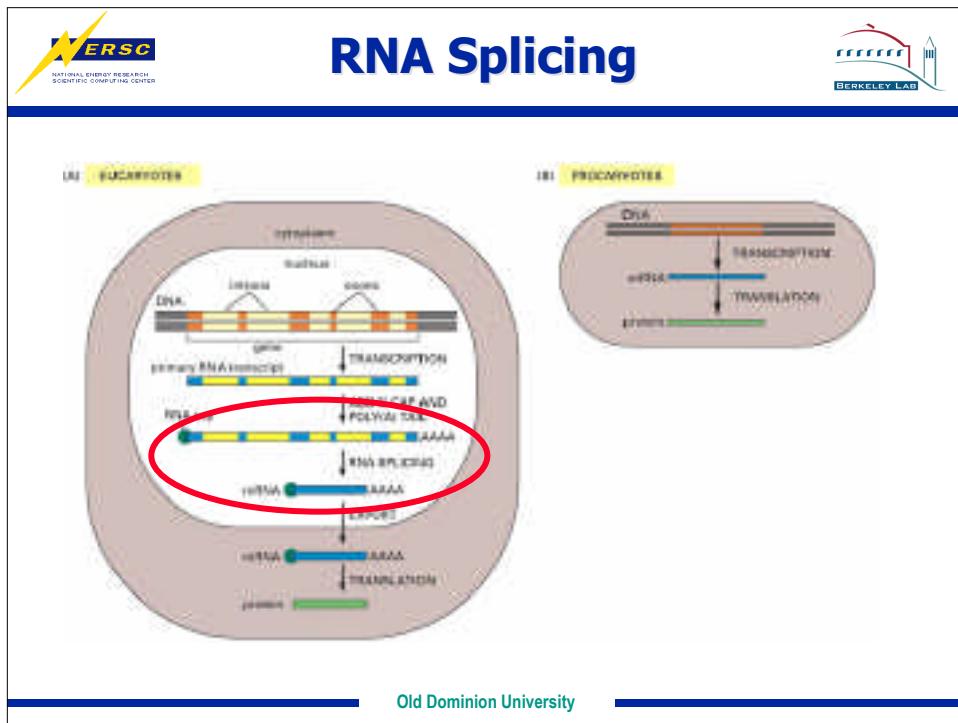
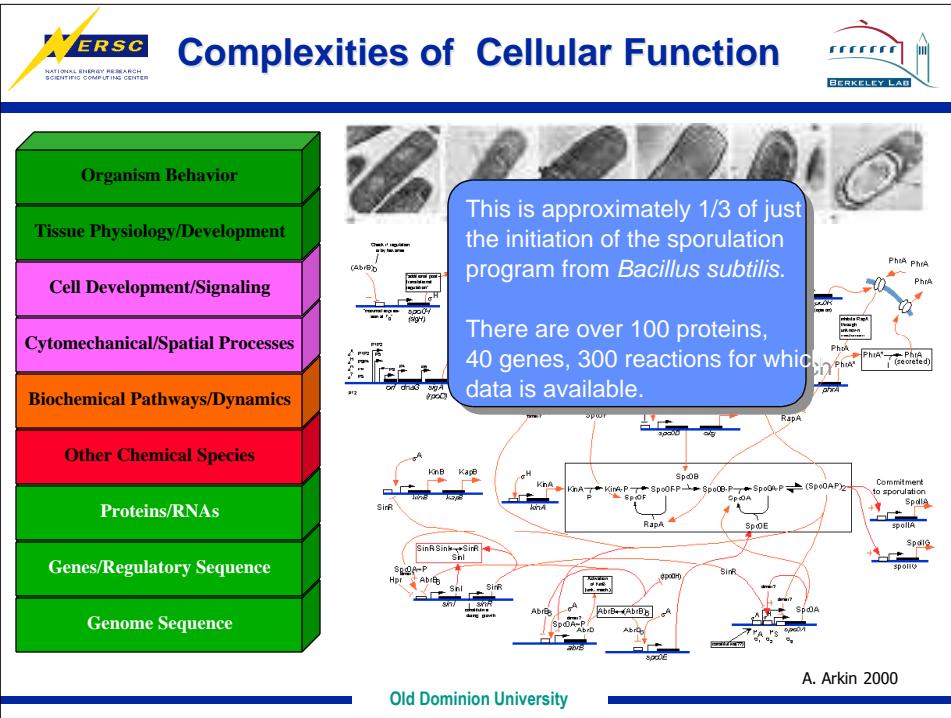
Layers of Information



**The same base sequence contains
many layered instructions!**

- **Chromosome structure and function**
 - Telomeres, centromers
- **Gene Regulatory information**
 - Enhancers, promoters, ...
- **Instructions for gene structure**
- **Instructions for protein**
- **Instructions for protein post-processing and localization**

Old Dominion University

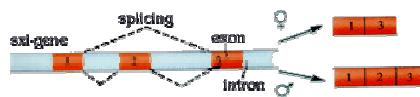




Gender in Drosophila



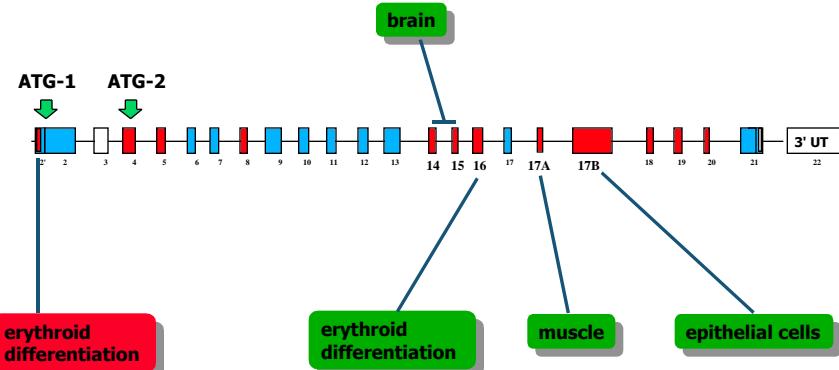
- A precursor-RNA may often be matured to mRNAs with alternative structures. An example where alternative splicing has a dramatic consequence is somatic sex determination in the fruit fly *Drosophila melanogaster*.
- In this system, the female-specific *sxl*-protein is a key regulator. It controls a cascade of alternative RNA splicing decisions that finally result in female flies.
- Sex in *Drosophila* is largely determined by alternative splicing



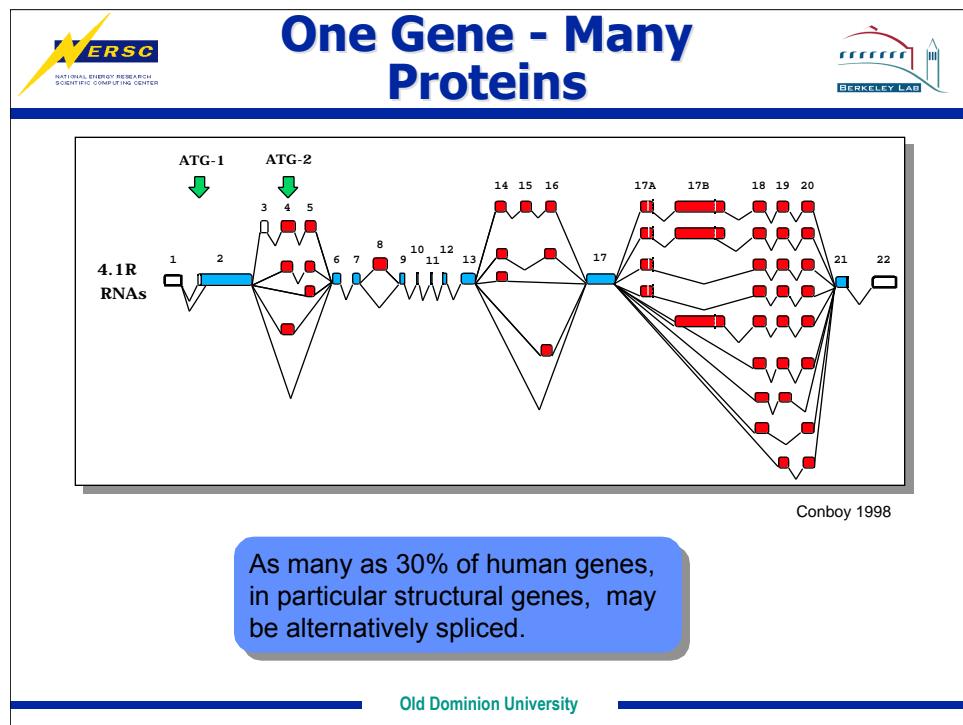
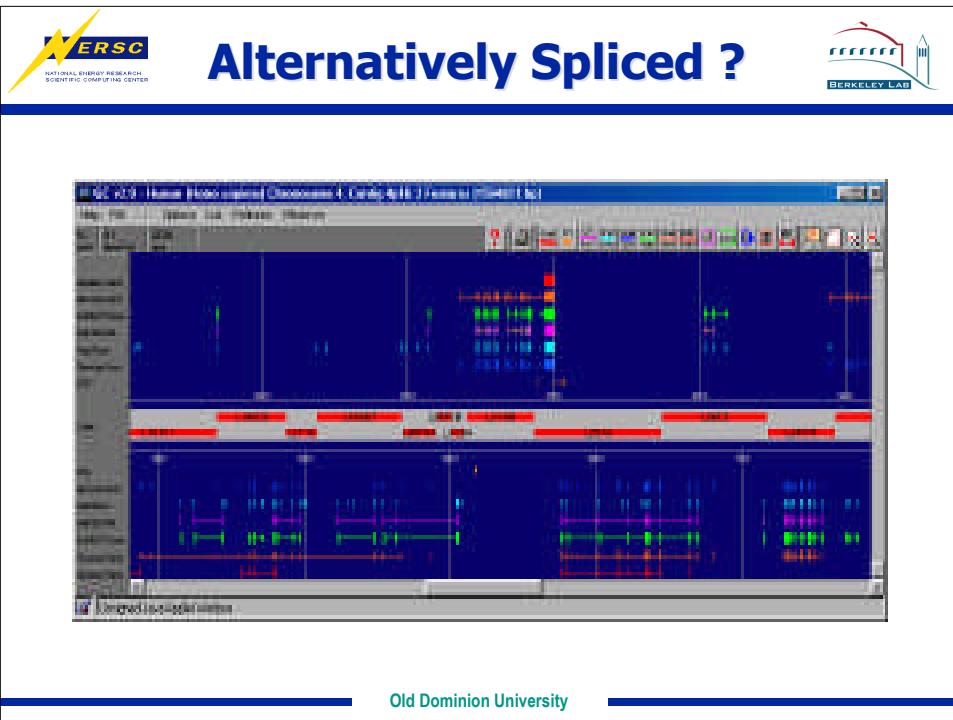
Old Dominion University



One Gene - Many Proteins

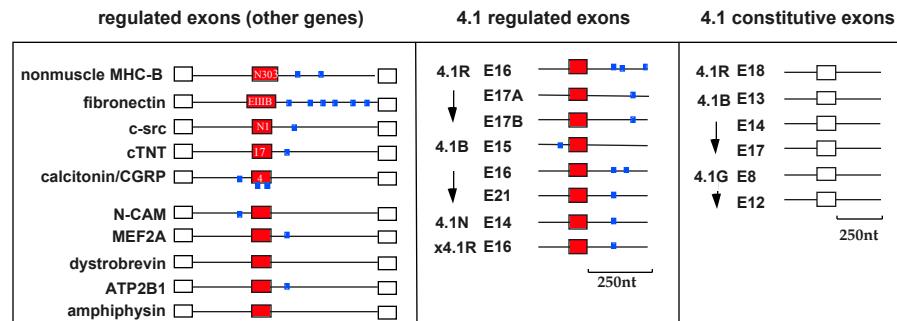


Old Dominion University





Regulatory Splice Signals



15 of 18 constitutive exons in ATP2B1;
15 of 18 unregulated exons in protein
4.1B lack UGCAUG elements

Old Dominion University

Alternative Splicing DB (ASDB) - Main Page - Netscape

File Edit View Quit Domains Help

View Insert Form Search Data File Help Home Reference

ASDB Home Page http://www1.lbl.gov/ASDB/

ASDB

Alternative Splicing DB

DB CONTENT | HOW TO USE | FURTHER WORK | SEARCH

References to the Alternative Splicing Database:

ASDB: database of alternatively spliced genes

J. Dralyuk, M. Brudno, M. S. Getz, M. Zom, and I. Dubchak (2000) Nucleic Acids Research 28(1), 295-297.

M. S. Getz, I. Dubchak, J. Dralyuk and M. Zom (1999) Nucleic Acids Research, 27(1), 301.

Alternative Splicing Database

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

BERKELEY LAB

Protein Prototype

- Selected 1,663 entries from SwissProt that contain the words "alternative splicing"
- 241 clusters of at least 20 aa overlap
- Multiple alignment

Search by Medline, SwissProt, GenBank identifiers; keywords, and feature tables

Old Dominion University

The screenshot shows a search result for 'DYT1 RAT' with 1,663 entries. A cluster information box highlights 'All members of the cluster' for C028_HUMAN, showing multiple sequence alignments. A blue box on the left details the protein prototype search results.

9p21 Gene Cluster is a Nexus of the Rb and p53 Pathways

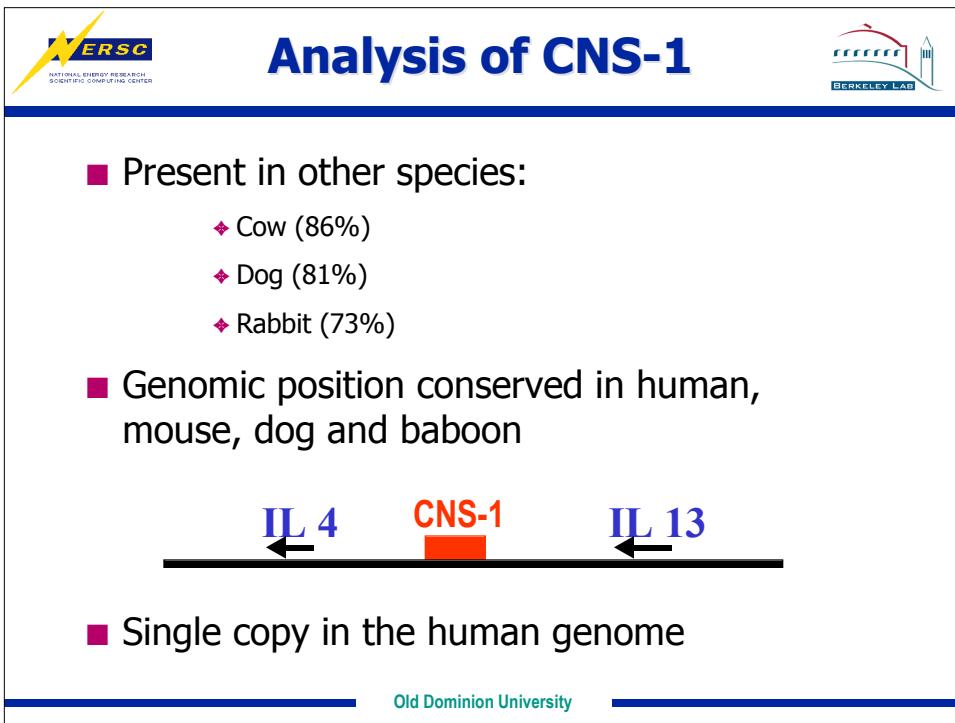
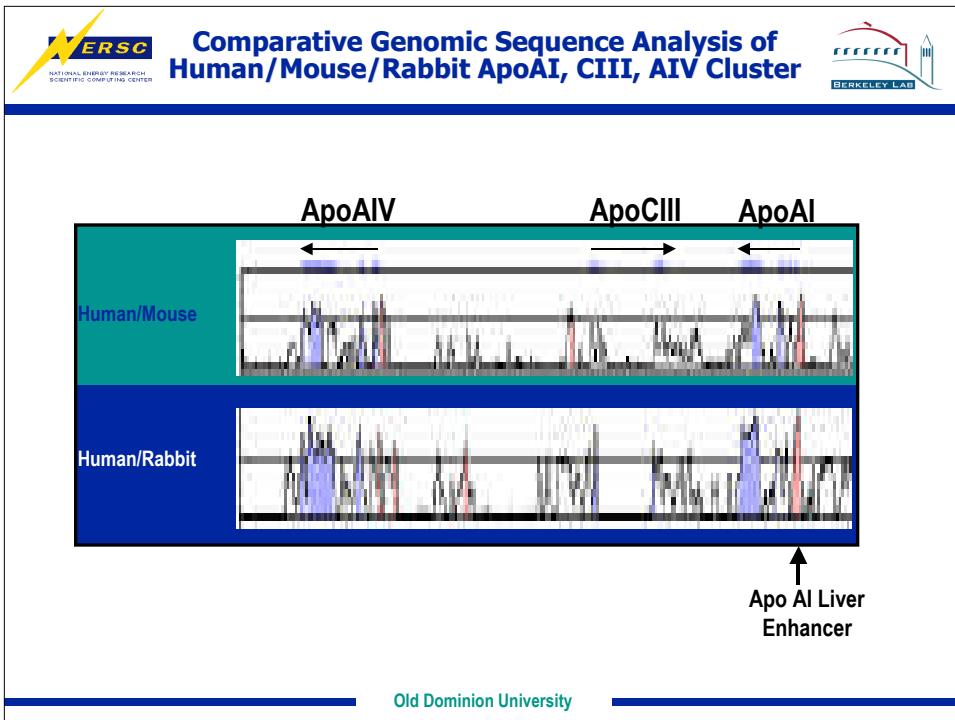
NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

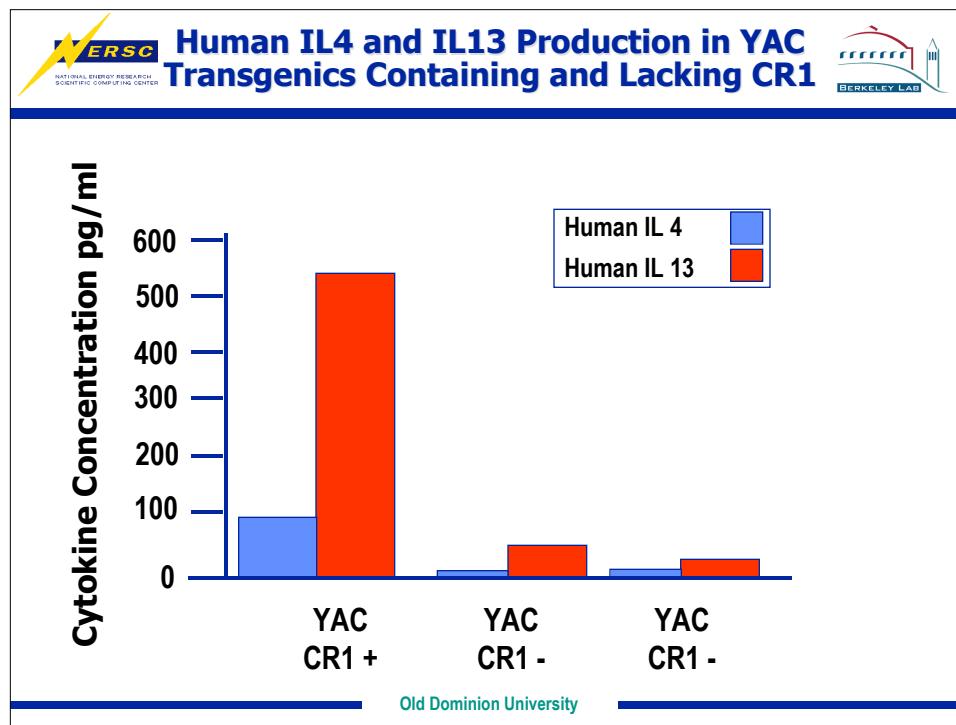
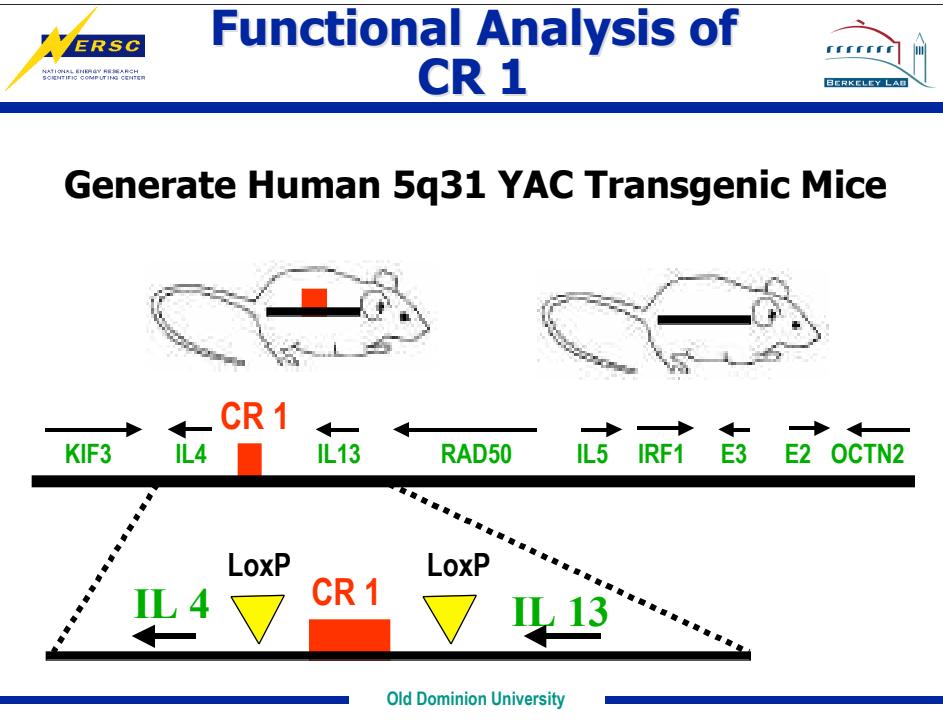
BERKELEY LAB

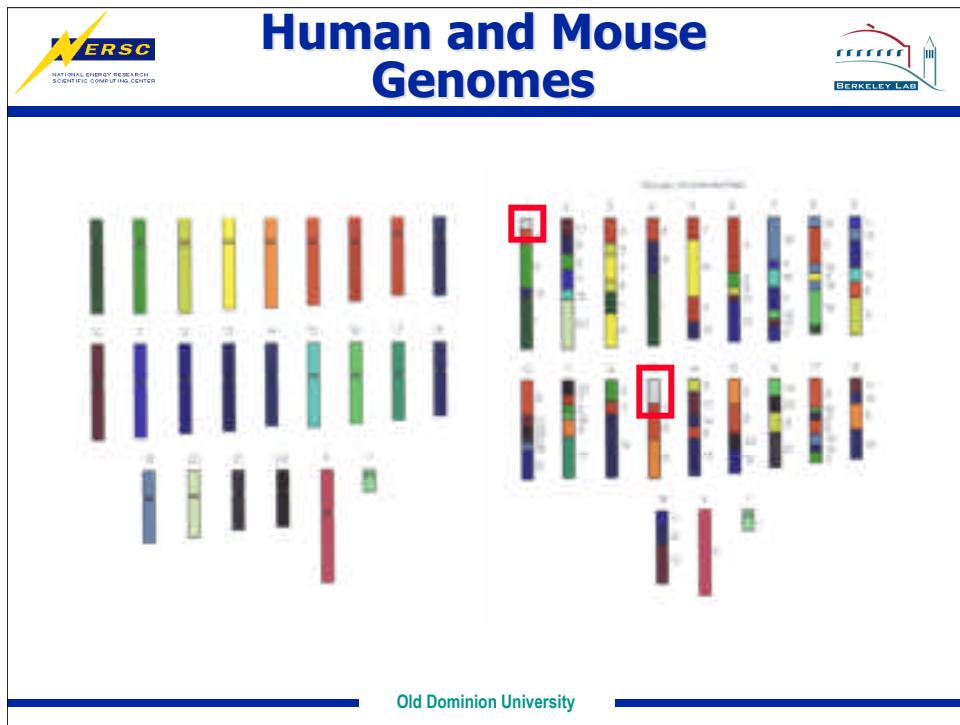
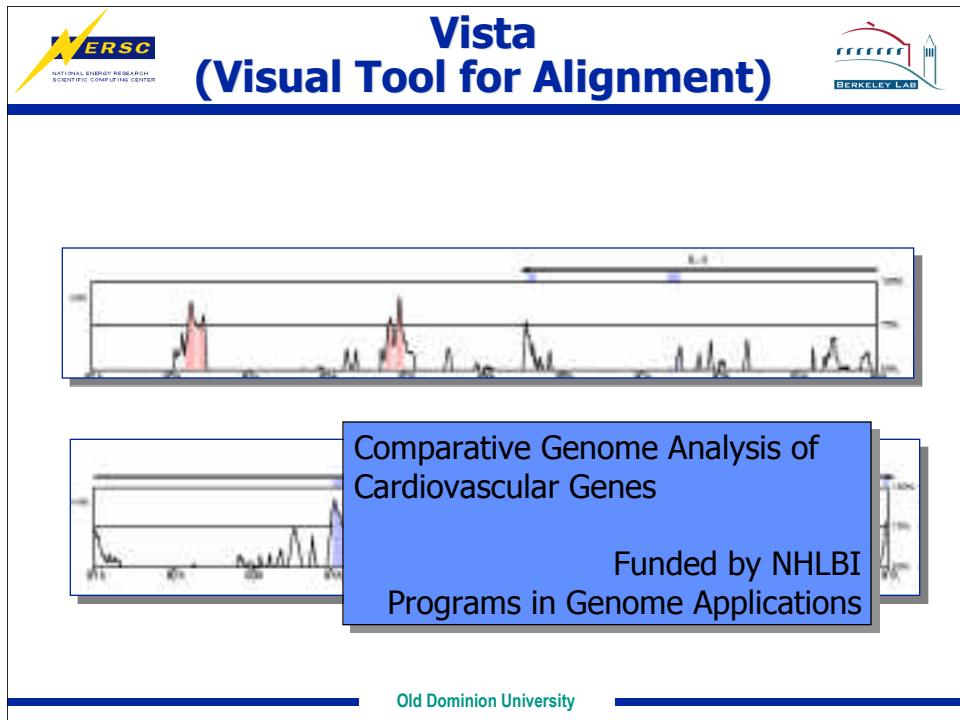
The diagram illustrates the 9p21 gene cluster with two alternative promoters (MTS2 bcr and MTS1 bcr) and three exons (E1, E2, E3). It shows how oncogenic stimuli (H-Ras) and extracellular stimuli (TGF- β) can induce different mRNA isoforms through alternative splicing. The p15^{INK4b} and p14^{ARF} genes are shown with their respective promoters and exons. A blue box highlights that both genes share the same partial nucleotide sequence but have different amino acid sequences due to alternative splicing.

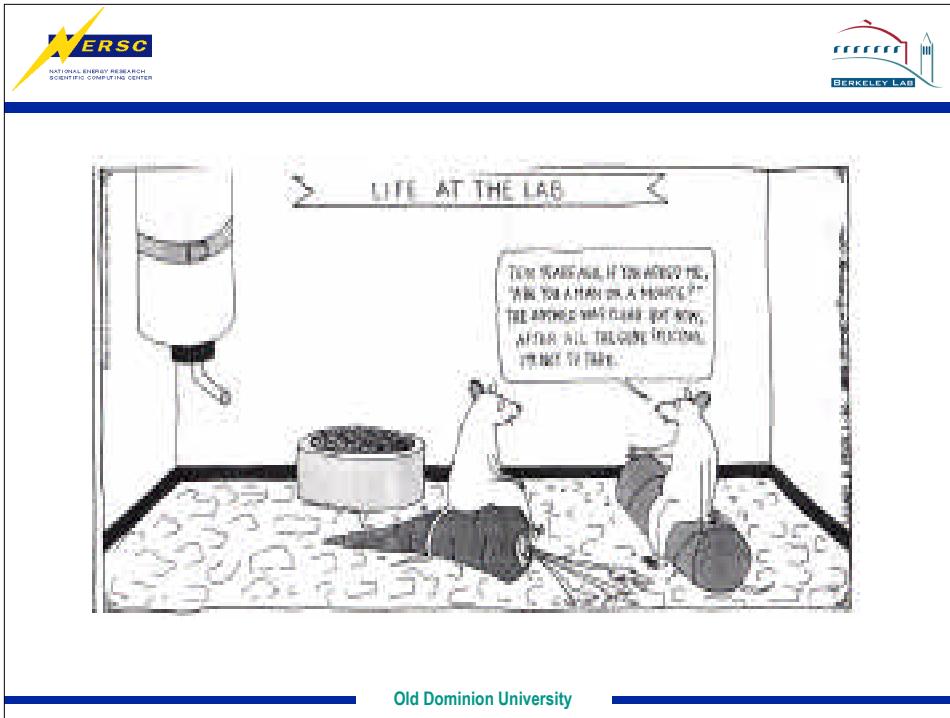
• Same partial nucleotide sequence
• Different amino acid sequence

Old Dominion University









Old Dominion University

"Genomical" Divide

- Traditional biology vs. Genomics
- Individual investigators vs. large, interdisciplinary teams

- Develop large data centers:
 - ✓ Data management
 - ✓ Data analysis
- Collaborative tool development
 - ✓ Lack of software integration
- High-performance computing
 - ✓ Beginning to build high-performance applications
- Global collaborations
 - ✓ Shortage of interdisciplinary scientists

Old Dominion University



What's Really Next



The post-genome era in biological research will take for granted ready access to huge amounts of genomic data.

The challenge will be *understanding* those data and using the understanding to solve real-world problems...

The end of the beginning rather than the beginning of the end.

Old Dominion University



Credits



■ NERSC / LBNL

- John Conboy
- Donn Davy
- Inna Dubchak
- Kelly Frazer
- Eddy Rubin
- Sylvia Spengler
- Eric P. Xing
- Manfred Zorn

■ ORNL

- Ed Uberbacher
- Richard Mural
- Phil LoCascio
- Sergey Petrov
- Manesh Shah
- Morey Parang

Old Dominion University